THE TIME TRAP: WHY IT'S MISGUIDED TO REPORT STATE ASSESSMENT RESULTS AS "YEARS OF LEARNING"

August 2024

Damian Betebenner
Center for Assessment

Charles A. DePascale
Psychometric Confections

Center for Assessment

National Center for the Improvement
of Educational Assessment
Dover, New Hampshire

Presented at the 2024 Annual Meeting of the National Council on Measurement in Education, Philadelphia, Pennsylvania.

The National Center for the Improvement of Educational Assessment, Inc. (the Center for Assessment) is a New Hampshire based not-for-profit (501(c)(3)) corporation. Founded in September 1998, the Center's mission is to improve student learning by partnering with educational leaders to advance effective practices and policies in support of high-quality assessment and accountability systems. The Center for Assessment does this by providing services directly to states, school districts, and partner organizations to support state and district assessment and accountability systems.

Authors are listed alphabetically and contributed equally to the ideas expressed in this work.

Betebenner, D. & DePascale, C. (2024). *The time trap: Why it's misguided to report state assessment results as "years of learning."* Dover, NH: The National Center for the Improvement of Educational Assessment, Inc.

TABLE OF CONTENTS

INTRODUCTION

The long-dormant practice of reporting student academic performance, particularly test results, in terms of weeks, months or years of learning has experienced a resurgence since the onset of the pandemic. The desire to communicate the magnitude of learning loss and recovery in an understandable way to non-technical audiences has made this practice commonplace. Time is an obvious choice as a metric, since PK-12 education is routinely broken down into units of time, and virtually all stakeholders (policymakers, educators, students, parents) ask how long it will take students to recover any learning lost during the pandemic.

The makers of widely used assessments were among the first to characterize the pandemic's impact on student learning. In May 2020, only a few months into the pandemic, states were determining whether schools should try to teach new content remotely or simply reinforce previously taught material when NWEA published its estimate that "students are likely to return in fall 2020 with approximately 63-68% of the learning gains in reading relative to a typical school year and with 37-50% of the learning gains in math" (Kuhfeld et al., 2020). In fall 2020, Curriculum Associates reported an increase in the percentage of students reporting to school two or more grade levels behind in reading and mathematics (Curriculum Associates, 2020) and Renaissance reported that student achievement in reading was "on average, only a single percentile point below where it should have been in a normal school year," but "math achievement has been significantly more affected … falling on average seven percentile points" (Renaissance, 2020).

The authors of these initial reports grounded their descriptions of the effects of the pandemic on student achievement in reporting metrics closely tied to their respective assessment programs, such as amount of growth expected in a year, percentage of students performing at grade level, and percentile ranks.

It was not long, however, before descriptions of *learning* lost to the pandemic shifted to a more familiar metric; that is, time. Percentile point losses on Renaissance exams were "translated into terms of instructional time" such as "students in grades 5 and 6 were more than 12 weeks behind beginning-of-year expectations in math" (Renaissance, 2020). NWEA reports sparked headlines such as, "Students need over 4 months of extra learning to return to pre-pandemic math, reading achievement" (Merod, 2023). A study conducted by World Bank reported learning losses on average "equivalent to roughly one-half year's worth of learning" (Patrinos et al., 2022). Noting score declines on the National Assessment of Educational Progress (NAEP), a study by McKinsey & Company said that "students in 2022 were on average about 15 to 24 weeks behind in math and nine weeks behind in reading compared with 2019" (Bryant et al, 2023).

On its face, converting student performance into units of time appears reasonable. There is certainly no lack of time-bound data on student achievement. Districts across the country administer interim assessments three or four times per year. States administer annual tests in reading and mathematics to all students in grades 3 through 8 and once in high school. NAEP tests in reading and mathematics are administered to representative samples of students across the country every other year.

For cross-sectional NAEP results, we can easily report how the nation, a state, or school performed on a particular test at two (or more) points in time. For longitudinal results, vertical scales make it possible to report on a common scale how much student performance has changed across those time points. Psychometric and statistical techniques ranging from the simple to the highly

sophisticated can be applied so that even changes in scores from different tests can be interpreted on a common metric.

We can state, with a high degree of confidence, time-based results such as:

- Performance of 4th grade students on both the 2022 NAEP reading and mathematics tests dropped to its lowest point since 2003.

Results on state assessments support comparative statements such as:

- Statewide performance on the 2023 7th grade mathematics test was 20 points higher than in 2021, but still 20 points below performance in 2019, recovering half of the 40-point difference from pre-pandemic results.

If the state reports assessment results on a vertical scale, we are comfortable making statements such as:

- Results on the 2023 Grade 8 reading test show that student performance increased by 150 points from 7th to 8th grade, consistent with the average gain in performance seen between those two grades before the pandemic.
- The average score of 500 on the 2023 Grade 8 reading test is 150 points lower than the average Grade 8 score before the pandemic.

We can even state with some degree of confidence, but considerably less interpretability:

- The 8th grade average score of 500 in 2023 is equal to the average score of 6th grade students in 2019, before the pandemic.

Our problems begin when we attempt to move beyond those relatively straightforward descriptions of point-in-time comparisons and interpolate or extrapolate from performance at known points in time to intervals for which we have no empirical data, in an attempt to:

- declare that students are two years behind, a half-year behind, or six weeks behind, which are inevitably interpreted as …
- … predictions or estimations of how much time will be needed to complete the recovery; that is, to close any remaining gaps between current and pre-pandemic performance or established goals.

In short, our problems begin when we attempt, through test scores alone, to provide time-based, non-technical, seemingly easy-to-interpret *information* that appears to describe the state of learning loss in terms of how "behind" students are and how long it will take them to recover.

In short, our problems begin when we attempt, through test scores alone, to provide time-based, non-technical, seemingly easy-to-interpret *information* that appears to describe the state of learning loss in terms of how "behind" students are and how long it will take them to recover.

Kuhfeld et al. (2023) warned that despite its popularity, time-based reporting "has some major downsides" and identified four important questions to consider when using time-based estimates:

1. Which assessment (or set of assessments) should I use to determine "typical" learning rates?
2. Is the population of students used to calculate the typical growth estimates comparable to my sample population?
3. Do I need a benchmark for a specific grade/subject or am I trying to generalize across multiple grades/subjects?
4. Are my months of learning estimates plausible? Would I get a substantially different answer if I made different choices for questions 1 through 3? (p. 2)

Although we agree with the relevance of each of these questions, we believe that issues associated with time-based estimates are much more fundamental and that recent attempts to convert test scores (data) into years of learning (information) in order to communicate with stakeholders fall short in four critical areas: semantics, technical foundations, empirical validation (or lack thereof), and utility. In this paper, we address each of those four areas:

Semantics: We begin with a discussion of the use of two key concepts at the heart of estimates and reporting of weeks, months, and years of learning: learning and time.

Technical Foundations: In this section, we provide an overview of grade-equivalent scores and the historical foundations of the technical processes, procedures, and assumptions underpinning current conversions of test scores into time-based metrics such as years, months, or weeks of learning.

Empirical Validation: Building on Kuhfeld et al.'s question, "Are my months of learning estimates plausible?", in the third section we present empirical data based on growth norms which, at a minimum, provide much-needed context to the interpretation of statements about years of learning.

Utility: In the final section, we discuss three *a priori* fundamental flaws regarding the utility of converting test scores to time-based reporting metrics and discuss whether current reporting methods adequately address those flaws.

The common thread running through the four sections is the question of whether the use of time-based interpretations of test results is likely to bring about changes in educational policy that support or lead to increases in instructional effectiveness, with the ultimate goal of improving student learning.

SEMANTICS

You keep using that word. I do not think it means what you think it means. – "The Princess Bride"

If the primary goal of educational assessment is to clearly communicate valuable and useful data to inform key stakeholders, such as educators, students, parents and policymakers, then it should be self-evident that the words and images we use to share assessment results with those audiences is of critical importance. Test scores, by themselves, are simple statements of student achievement which convey very little information. However, the ways we present, describe, and explain them, and the words that we use when discussing test scores, often confuse people. There is ample evidence

to support the claim that historically, such communication is not the field of educational assessment's strong suit (Goodman et al, 2004; Zwick et al, 2014; Zenisky et al, 2015).

One area in which our field often falls short is in the imprecise or inappropriate use of terms when we're communicating test results among ourselves and with critical stakeholders. One such example is our use of the word *learning*, and the phrase *student learning*, as synonymous with *achievement*. One unintended consequence of conflating learning and achievement has been the expectation that a summative achievement test can provide actionable information to improve instruction and student learning. We feel that the current focus on time is another prime example of an unintended consequence and undesirable outcome associated with conflating student learning and achievement.

Learning: The unfortunate rise of a misleading label: "learning loss"

As the pandemic unfolded, there was significant consternation over the widespread use of the term "learning loss"; people argued about its meaning, its implications, and the mindset it created. At the time, the word "loss" was what generated most of the concern (Schwartz, 2021; UnboundEd, 2021; Whitby et al, 2021):

- Did learning *loss* imply that students actually know less in September 2020, June 2021, or September 2021 than they did when COVID shut down schools in March 2020?
- Did the use of the word *loss* unfairly place the onus for the current "state of learning" on the students who had "lost" something during the pandemic and/or the teachers who struggled to facilitate student learning in a pandemic?
- Did the use of the word *loss* feed into the deficit mindset already prevalent in public education before the pandemic?

These are all good and important questions, but there should have been at least as much attention devoted to the use of the word "learning." As scores from various tests—first commercial interim assessments, then state tests, and eventually NAEP—became the primary indicators (some would say measures) of learning loss, and then learning recovery, it did not take long for scores on those tests to become synonymous with student learning.

The problem is that none of those tests measure *student learning.* That's because no single test can measure student learning. Learning is not **an outcome** that can be measured on a single test—let alone an achievement test. Learning is **the process** that produces a change in student achievement between the administration of two tests measuring that achievement. Student learning is not student achievement. Fair questions that more accurately reflect the relationship between student achievement and student learning might have been:

- What factors during the pandemic led to lower achievement than expected/desired at key academic milestones (e.g., the end of the school year or the time of testing)?

Learning is not **an outcome** that can be measured on a single test—let alone an achievement test. Learning is **the process** that produces a change in student achievement between the administration of two tests measuring that achievement.

- To what extent was lower achievement related to factors such as the amount or quality of instruction, and to what extent was it caused by factors such as stress, which could have undermined students' ability to process information provided during instruction as effectively?

Expressed in terms of student learning rather than simply achievement, the same questions might have identified factors that led to a decrease in the *rate* of student learning (the change in student achievement over a fixed period of time) during the pandemic. We will address the relationship between learning and achievement in more depth in subsequent sections of the paper.

The current misuse of the term "learning" is particularly disappointing because we have been down this road before—more than once, and fairly recently. In the early 2000s when the assessment requirements of the No Child Left Behind Act were fully implemented, we mistakenly conflated student achievement and school effectiveness (Popham, 1999). In the 2010s, we mistook student achievement (in the form of test scores) as a legitimate indicator of educator effectiveness (Baker et al, 2010). Today's conflation is student achievement and student learning. By making the same mistake over and over, we've become, in the parlance of former New England Patriots coach Bill Belichick, *error repeaters*. And repeating errors is a strong indicator of a lack of learning.

It would be easy to dismiss the concern about the misuse of the term learning as nitpicking, or as being pedantic, if the practice did not have consequences for interpretations and actions. As a starting point, an unintended consequence of conflating learning and achievement is that it results in the same type of misinterpretations previously associated with school and educator effectiveness. A low score on an achievement test (i.e., a status score) is no more a first-degree indicator of student learning than it is of school effectiveness or teacher/teaching effectiveness. Treating student achievement as a first-degree indicator of school effectiveness made it less likely for policymakers to address known underlying causes of differences in student achievement such as inequity in funding and resources, access to quality instruction and materials, and a safe and secure learning environment. Similarly, conflating student achievement and student learning makes it likely that we will continue to assess the outcome of student achievement while giving short shrift to the assessment/measurement of the input/process factors that have long been established as indicators of student learning, such as time on task, active responding, feedback, questioning, and engagement.

Time: Out of favor as a determinant of achievement

Although most of our concerns about the focus on time fall into the other three categories addressed in this paper (i.e., technical, validity, utility), we would be remiss not to state that we are befuddled by the semantic choice to use time as the vehicle to interpret test scores precisely when the field has taken a hard turn away from using time as a relevant factor in evaluating or determining student achievement.

At the macro level, the decades-long pushback against the concept of seat time has prompted the Carnegie Foundation, in partnership with ETS, to lead the effort to move beyond the Carnegie Unit—the primary time-centered unit applied in preK-12 education, particularly secondary education, for more than a century (Carnegie Foundation, 2023). At the micro or individual student level, a focus on time is the antithesis of the competency-based movement in instruction and assessment. The following quote from a conference presentation is but one manifestation of that movement's position on time (Wormeli, 2022):

> *Time is not immutable. It's a variable. Popcorn kernels pop at different rates, but when each one pops, it's accorded full status as a piece of popcorn, not something less than popcorn because it popped later than its fellow kernels. We are not beholden to an arbitrary, uniform timeline.*

At a minimum, the choice to emphasize time in the interpretation of test scores seems out-of-touch with the rest of education. This isn't, unfortunately, unfamiliar to educational assessment, but our field has been making a concerted effort in the past decade to stop being so out of touch with the world and the people we serve.

In addition to demonstrating the distance between time-based reporting and the center of current thinking in preK-12 education, the popcorn quote also serves as a nice transition into the next sections on our technical, accuracy, and utility concerns about reporting test scores in terms of time. That's because time, in fact, is immutable, inexorable, and a central factor in student learning. No matter how much we wish that were not the case, time marches on; it waits for no one. What is not immutable is the relationship between time and student achievement; that is, the rate of student learning.

TECHNICAL FOUNDATIONS

Measurement is never better than the empirical operations by which it is carried out, and operations range from bad to good. – S.S. Stevens (1946)

In this section, we discuss the technical processes and procedures which have been used historically to make the connection between test scores and time, the assumptions that are made regarding student achievement over time, and their implications and ramifications for the current interpretation and use of time-based metrics to report test scores.

Perhaps the most well-known *transformed score* relating student achievement to a particular point in time is the *grade equivalent* (GE) score, a norm-referenced score that was widely reported on standardized norm-referenced tests. Placing GE in the context of our previous discussion about the primary purpose of assessment (communicating valuable information), GE represent an attempt to convert data (in this case, test scores) into information (grade level context) as a primary means of communicating test results to stakeholders.

The GE score is defined as the median score attained by students at a particular grade level taking a test at a particular point in time (i.e., a particular month during the school year). It typically ranges from 0.0 to 12.9, depicting performance from the beginning of kindergarten through the end of 12th grade. For example, on a Grade 6 reading test, the median score attained by 6th grade students taking the test at the very beginning of 6th grade would be converted to a GE score of 6.0. Similarly, the median score attained by 6th grade students taking the same Grade 6 test, or a parallel form of it, at the end of the typical 10-month school year would be converted to a GE score of 6.9 (Kern, 2023).

So far, this seems straightforward and easy to interpret. The interpretation of GE scores becomes a bit cloudier when they extend beyond the grade level of the test, as a sizable percentage of scores in any given year will do. A 6th grade student taking the Grade 6 reading test, for example, might receive a GE score of 8.3. Educators are told that the "correct" interpretation of the GE of 8.3 is that our 6th grade student received the same score as the typical (median) student in the third month of 8th grade taking the Grade 6th grade reading test. They are warned that a GE of 8.3 does not mean that the 6th grade student has mastered 8th grade content. Still relatively easy to interpret after the explanation is provided, but the usefulness of the information is somewhat limited.

Interpretation becomes much more convoluted for students receiving a GE below their current grade level. (Note that on a test administered at the beginning of the school year, by definition, half of the students will score below the median, hence receiving a GE below their current grade level.) A 6th grade student taking the Grade 6 reading test at the beginning of 6th grade and scoring below the median might receive a GE score of 5.7, 5.2, or 4.9, dependent upon just how far below the median they score.

Interpretative materials prepared for educators and parents, however, rarely explain a GE score of 4.9 on the Grade 6 reading test in a manner parallel to that of GE scores above grade level. It is not described in relation to the Grade 6 reading test as the score that a typical student at the end of 4th grade would score if they took the Grade 6 reading test. Neither would they typically describe the GE of 4.9 as a score that our low-performing 6th grade student would have scored on the Grade 4 reading test. People don't discuss low GE scores in either of those ways for good reason; neither is easy to interpret. This lack of interpretability of low scores is particularly troubling given that, as in the case of pandemic recovery, we are usually more concerned with low test scores than with high test scores.

At this point, you may be wondering why we are spending so much time discussing GE scores when a) they were soundly rejected years ago (Berk, 1981; Bennett, 1982; Ramos, 1996) and b) they are not what is being used today when testing programs and researchers report that students are one year, two years, or a half-year behind. The answer is that although some of the data analytic techniques used today have changed, the principles underlying the time-based estimates made today and the estimation of GE scores in the past are the same, and GE scores demonstrate the problems associated with attempting to convert test scores and changes in test scores to units of time.

Principles and problems associated with GE scores

GE Scores Are Not Numbers

As mentioned above, GE scores are presented as numbers ranging from 0.0 to 12.9 and are often depicted on a number line. The problem is that a GE of 3.4 or 5.7 or 6.9 is not a number in the conventional way that we regard numbers in measurement. That is, a GE score is not a cardinal number depicting a quantity of something, in this case achievement at a point in time. A GE is an ordinal number, or more accurately, a combination of two ordinal numbers representing grade level and month. A clue to the ordinal nature is found in the way that we read a GE score, for example, "fourth month of third grade" (3.4) or "seventh month of fifth grade" (5.7). First, second, third, etc., are ordinal statements indicating rank order, not cardinal statements indicating quantity. The symbol "." between the two ordinal numbers in a GE score such as 5.7 is simply a period or dot, not a decimal point. It could just as easily be replaced by a hyphen (5-7), slash (5/7), or set of brackets (5[7]) to convey the same meaning.

The first bit of deception in how we treat GE scores is in the use of 0. Zero is not usually associated with ordinal numbers. By reporting GE scores as cardinal numbers (and ultimately as real numbers) on a number line, we are implying scale properties that likely cannot be supported, particularly with regard to the performance of an individual student. Similarly, the use of time-based language (years and months) implies the existence of an underlying ratio scale (i.e., time or duration) or at the very least, an interval scale.

The Empirical Illusion

As described above, the definition of GE scores is grounded in empirical descriptions. A GE of 6.7 on the Grade 6 reading test represents the score that a typical 6th grade student would receive when taking the test in the seventh month of the school year. A GE of 8.3 on the Grade 6 reading test represents the score that a typical 8th grade student in the third month of the year would receive on that test. Unfortunately, those GE scores typically are not based on performance of samples of actual 6th or 8th grade students taking the Grade 6 reading test at those particular points in time. The scores are derived from statistical analyses; that is, extrapolations and interpolations based on certain assumptions about the distribution of scores on different tests taken at different grade levels. So, how is it that we moved from an empirically based ordinal number (the median score of an actual sample of students taking a test at the beginning or end of a grade level) to GE based on distributions of test scores across grade levels and tests?

Now Entering: Vertical Scales

To a large extent, GE scores and the estimates of years of learning that we see reported today are dependent upon vertical scales; that is, a common scale linking scores on two or more different tests. With vertical scales, it is not necessary to administer the same test to 4th, 6th, and 8th grades students to obtain scores on a common scale. A high-achieving 4th grade student taking the Grade 4 test, an average-achieving 6th grade student taking a Grade 6 test, and a low-achieving 8th grade student taking the Grade 8 test can all be assigned a score on the same scale; and in fact, they may all be assigned the same scaled score.

Still, the connection between test scores, grade levels, and time is not as direct as might be implied by the existence of a common scale. It is not as simple as assigning a single time-based score to a particular scaled score; for example, using the median or mode to assign a grade-based estimate to every score on the vertical scale. In practice, the estimates that formed the basis for GE scores and time-based metrics used today flow from a single piece of information on the vertical scale: the change in student performance between two annual test administrations.

The Single Seed: A Year's Worth of "Learning"

In general, the key piece of empirical data that we have in creating time-based estimates is the change in the achievement of students on annual tests administered at the end of successive grade levels. For example, we know students' average scaled score on a Grade 4 mathematics test and on a Grade 5 mathematics test. (Note: These may be the same students in successive years or "randomly equivalent" cohorts of students in a single year.) Those average scaled scores on each test-based scale are converted to scores on the common vertical scale. The difference between those scores represents a *year's worth of learning*, or, more accurately, the c*hange in student achievement* from the end of 4th grade to the end of 5th grade. From that single figure, the rest of the calculations of time-based estimates blossom.

Interpolations, Extrapolations, and Standard Deviations

With a figure in place to define a year's worth of learning, analyses return to the individual grade-level tests and the distributions of scores on those tests. In simplistic but fairly accurate terms, time-based estimates are created by applying that value for a year's worth of learning to the distribution of scores on each grade level test. For example, if a *year's worth of learning* is equal to 100 scaled score points and the standard deviation of scores on the Grade 5 reading test is 100 scaled score points, one standard deviation represents a year's worth of learning. Accordingly, 150 scaled score points (1.5 standard deviations) represent 1.5 years of learning; 25 scaled score points (0.25 standard deviations) represent a quarter-year of learning; and so on. There is no attempt to connect these time-based estimates to the performance of a hypothetical student on a particular grade-level test. There is no attempt to connect these time-based estimates to the knowledge and skills that such a student might possess.

Assumptions, Choices, and Distributions

It is important to note that no matter how meticulously built these grade-level and vertical scales are, they are based on assumptions and choices related to the distributions of scores on each grade level test and in the construction of the vertical scale. Different, equally valid choices in the construction of the vertical scale will result in different estimates of a year's worth of learning. Different, equally valid choices in the design of the grade level test, the sample of students tested, and the conversion of test scores to a grade-level scale will result in different standard deviations and different relationships between test scores and years of learning.

Takeaway points

The statistical machinations used to generate time-based estimates are technically sound, but they are based on series of assumptions, and a small bit of empirical data. The time-based conversions may be internally consistent within the system in which they were created, but may not translate well outside of that system in a way that supports the design of interventions to improve student learning, and ultimately, student achievement. By the time we move from raw test scores to scaled scores, to average scaled scores, to vertical scale scores, to differences in vertical scale scores, to standard deviations, to effect sizes, and finally to time-based descriptions of scores, we are far removed from the actual test and its contents, and far from identifying gaps in content knowledge and skills, or determining whether or how those gaps should be filled, and how long it might take to fill them.

Further, whatever the statistical nature of the intervals reflected by the scores on a vertical scale, they are no more "equal" than the months of a year in terms of the rate of learning and the amount of achievement that can be expected between them. When directly comparing vertical scale scores obtained on different tests, it is difficult to argue that those *scaled scores* are numbers any more than GE scores were. The "number" represents relative position, but tells little more about what a student, or group of students, knows and can do. Converting scaled score change to units of time, however, implies the existence of properties of scaled scores that arguably don't exist. Time is on a ratio scale, and scaled scores are on an interval scale (at best). Growth over time is on neither.

Converting scaled score change to units of time implies the existence of properties of scaled scores that arguably don't exist.

Ironically, perhaps, given all of their flaws, GE scores still retain their "ordinal" nature in the way that they are presented, so

they may be less susceptible to misinterpretation than current time-based reporting expressed solely in terms of years. First, unlike current time-based reporting, GE scores were always grounded within a particular grade level. Second, people receiving GE scores who are familiar with schools, such as administrators, teachers, parents and students, are well aware that learning is not distributed evenly across the school year. The amount of learning that occurs between the beginning of the school year and Thanksgiving is different than the amount of learning that takes place between New Year's and early spring, and different still than the amount of learning in the final month or two of the school year. Learning is not linear.

EMPIRICAL VALIDATION

"Time is an illusion." – Albert Einstein

Given the administration of vertically scaled tests in numerous states, it is possible to empirically examine the extent to which assessment results are amenable to conversion into "years of learning." Vertically scaled tests allow for the comparison of scores across adjacent grades given the manner in which IRT scaling is performed.

In what follows, we use vertically scaled assessment results from multiple states to examine the frequency with which we empirically observe "two years' worth of learning" based upon observation of students across two years. We then examine how frequently two years' worth of learning occurs in a single year. We also observe the frequency with which students make "zero years' worth of learning," based upon the assumption that no change in their scaled score across a single year is equivalent to no learning having occurred. Finally, for completeness' sake, we also observe the frequency with which we observe students making three and four years' worth of learning in a single year.[1]

Following Yen (1986) and Dadey and Briggs (2012), cross-grade effect size is defined as the standardized mean difference between grade two content area means ($\bar{X}_{g1}$) and ($\bar{X}_{g2}$) using the pooled standard deviation:

$$Effect\ Size \equiv \frac{\bar{X}_{g1} - \bar{X}_{g2}}{\sqrt{\frac{\sigma_{g1}^2 + \sigma_{g2}^2}{2}}}$$

[1] We recognize that score comparisons of vertical scale scores across multiple years are problematic given the way that vertical scales are constructed via item overlap in scale construction. We include these calculations for completeness' sake, recognizing that the zero and two-year learning spans are most relevant.

There are several important aspects of this definition worth noting:

- The effect size associated with a year's worth of learning is standardized, average distance; that is, it's largely a normative definition.
- To calculate this effect size requires that the scores in the numerator be on the same scale so that subtraction makes sense; that is, we require a vertical scale.
- The effect size quantifies the extent to which the mean has shifted between the two scaled score distributions.
- There will be considerable overlap in the students in g_1 and g_2 in most applications of the above equation to state summative assessment data.
- However, there is no requirement that g_1 and g_2 consist of the same students (e.g., NAEP).

Given the way the "years of learning" idea has permeated discussions about individual student learning loss, it is instructive to consider this definition from the perspective of longitudinal student data. Figure 1 is a scatterplot of Grade 3 versus Grade 4 individual level student scores in mathematics. The effect size change (a year's worth of learning for these students) is 0.41, shown on the right vertical axis. The x- and y-axes provide both scaled scores and standardized scaled scores based upon the pooled standard deviation indicated in the definition above.

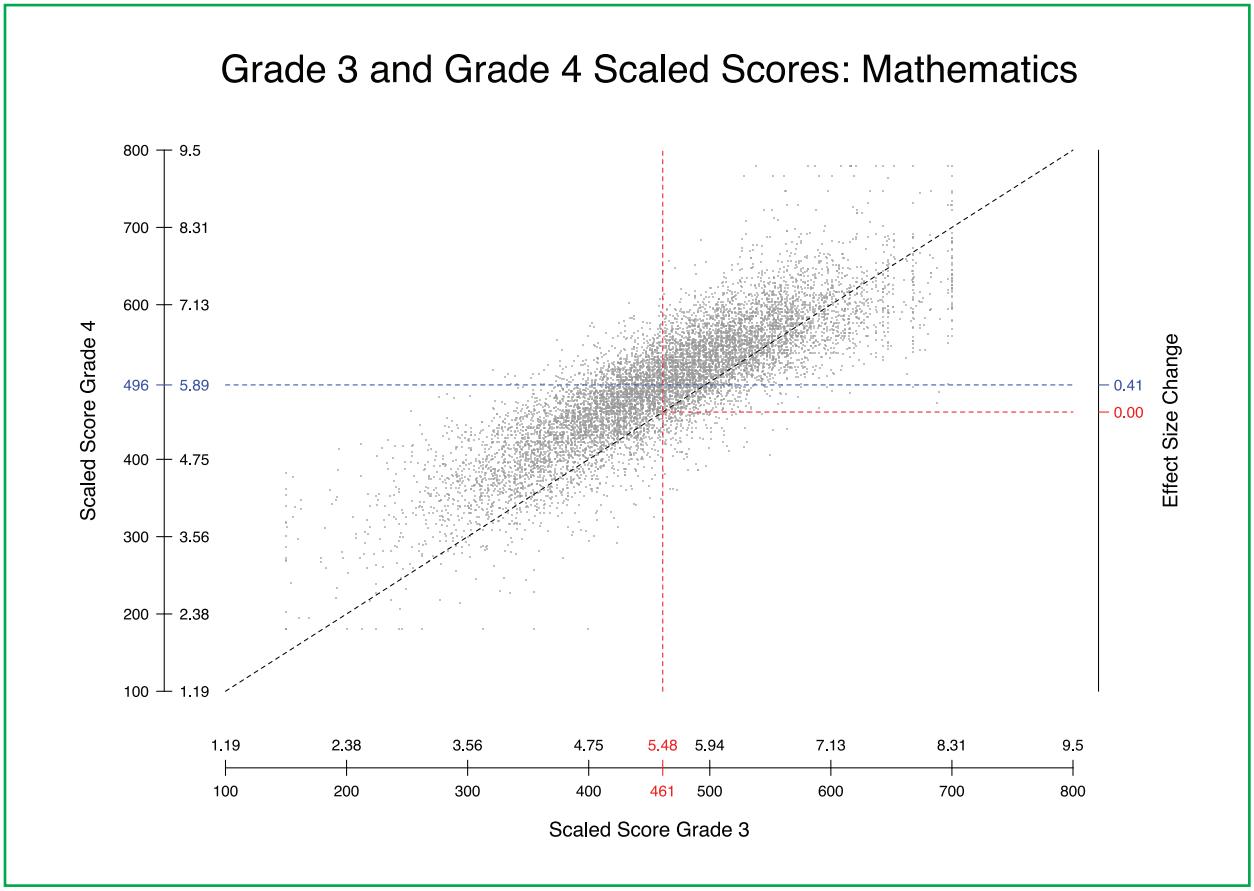


Figure 1: Scatterplot of grade 3 versus grade 4 individual student mathematics scores .

The figure illustrates the effect size gain associated with a year's worth of learning as a blue-dashed horizontal line. The standardized mean score for grade 3 equals 5.48, whereas for grade 4 it equals 5.89. The difference of 0.41 equals the effect size increase (a year's worth of learning). The identity line is also provided showing where grade 3 and grade 4 scores are identical.

The effect size definition for a year's worth of learning is an aggregate indicator that isn't suitable for individual-level determinations of whether a student attains or doesn't attain a year's worth of learning. To see this, Figure 2 illustrates two regions from the scatterplot of Figure 1 with the difference between scores (i.e., grade 4 – grade 3) either greater than or less than the average effect size difference. In addition, the bottom axis provides percentages of students whose gain is greater than the effect size (0.41) by grade 3 achievement decile. In the lowest grade 3 achievement decile, 74.4% of students demonstrate gains larger than the average effect size, whereas in the highest decile only 15.7% of students demonstrate a gain larger than the effect size.

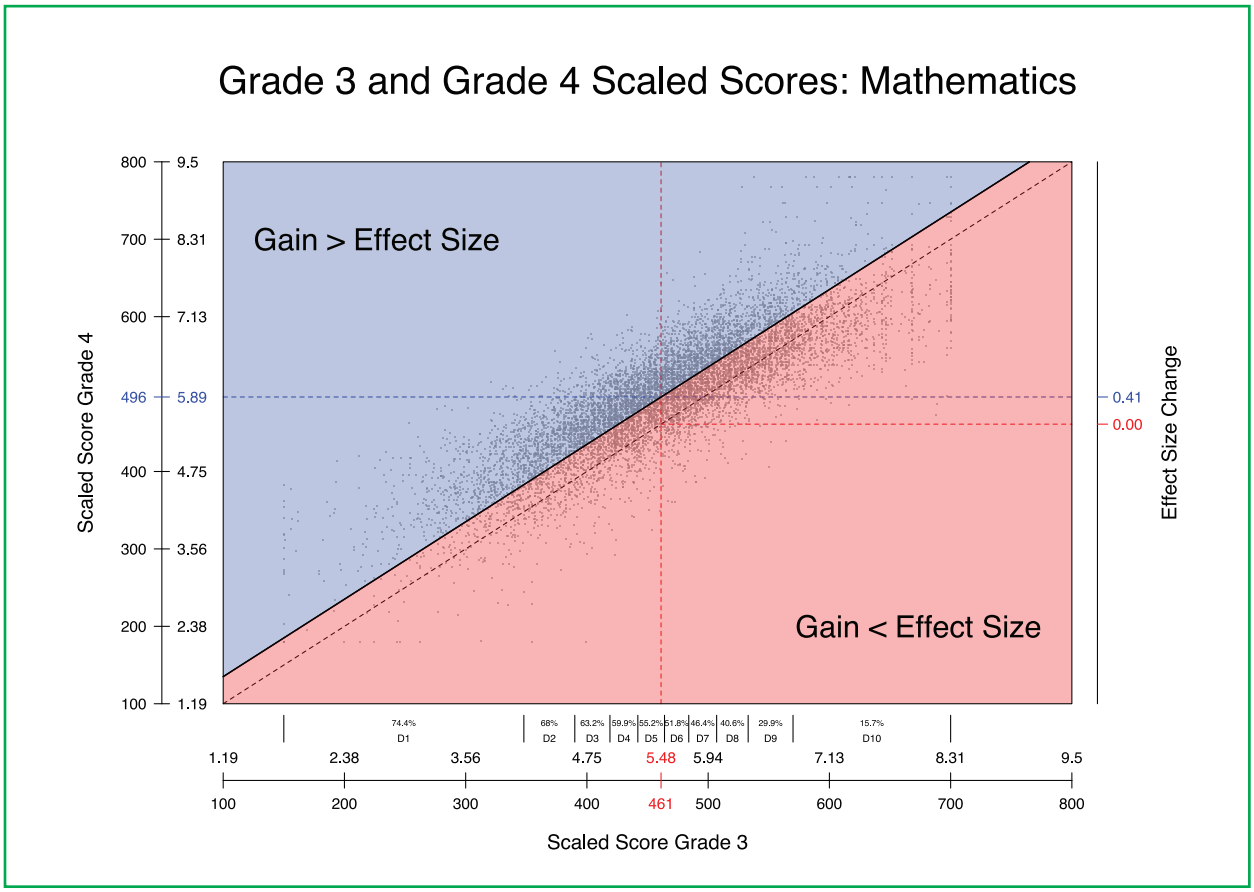


Figure 2: Scatterplot of grade 3 versus grade 4 student mathematics scores with regions indicating gains > effect size and gains < effect size.

This difference in gains between the highest and lowest achieving students, of course, reflects the well-known negative correlation between gain scores and prior achievement. To extend the notion of a year's worth of learning in a way that is consistent with the effect size gain, one needs to realize only that the point $(\bar{X}_{g1}, \bar{X}_{g2})$ lies both on the line indicating the effect size gain associated with grade 4 - grade 3 scale scores as well as the line of grade 4 scores regressed on grade 3 scores (illustrated

in Figure 3). That is, the normative nature of the "year's worth of learning" definition based upon an effect size difference extends naturally using regression to define the threshold above/below which a student has been deemed to attain (or not attain) a year's worth of learning.

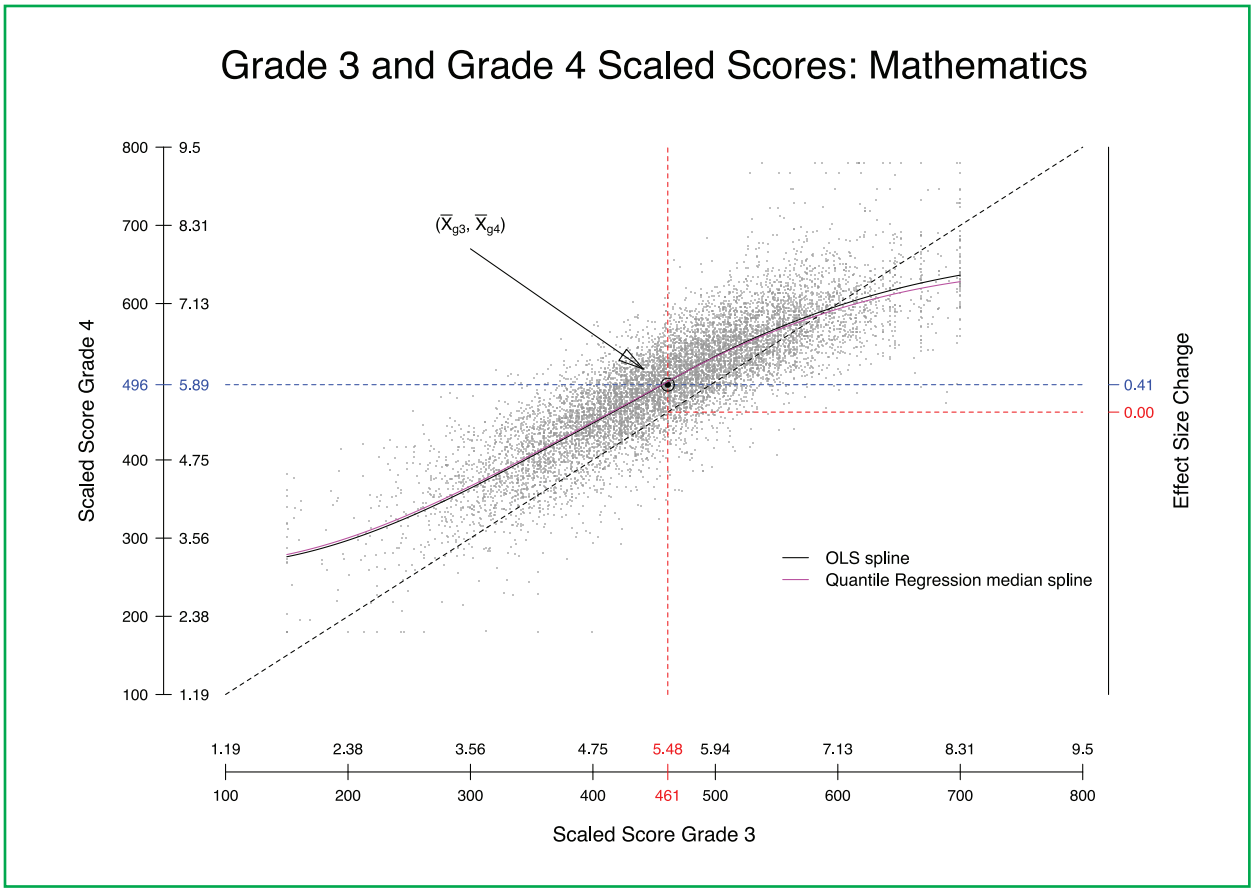


Figure 3: Scatterplot of grade 3 versus grade 4 mathematics scores with OLS spline and median regression spline overlaid on top.

This definition of a year's worth of learning comports well with the normative framework for individual student growth provided by student growth percentile (SGP). As illustrated in Figure 3, the mean and median regression spline are nearly identical, indicating than an individual student growth percentile (SGP) of 50 is the natural extension of the effect size definition to the individual student level.

Data and methodology used in the analyses

It is through this conceptualization that we investigate zero, one-half, two and three years' worth of learning as indicated in vertically scaled tests. Using vertically scaled state assessment data from four states, we calculate consecutive grade growth norms (coefficient matrices) using the Student Growth Percentile (SGP) framework for student-level calculations as well as growth norms extending across two years. Using these growth norms, we then leverage the purported equivalence of scores across grades and feed in data to those norms spanning zero, two and three years to determine the one-year SGP associated with zero, two and three years' worth of learning for each student.

- 0 years of growth: Assume identical scores for students from one grade to the next and feed those data into the one-year growth norms
- 2 years of growth: Use scores for individual students spanning two years (e.g., grade 3 and grade 5) and feed those data into the one-year growth norms.
- 3 years of growth: Use scores for individual students spanning three years (e.g., grade 3 and grade 6) and feed those data into the one-year growth norms.
- 1/2 year of growth: Use scores for individual students spanning one year (e.g., grade 3 and grade 4) and feed those data into the two-year growth norms.

Note that SGP, being a percentile, is a probability statement on the likelihood of such growth happening. By definition, one year of growth will equal 50th percentile growth. The question we investigate is the likelihood of observing zero, two and three years' worth of learning.

Results for a single state's data are presented in Figure 4. For this state, each grade consists of approximately 65,000 students in both English/language arts and mathematics. Results for zero, one-half, one, two and three years' worth of learning are represented in different colors. Grades 3, 4, 5, 6 and 7 are represented from left to right, and within each grade, the five different quintiles (Q1, Q2, Q3, Q4, and Q5) are laid out from left to right. Again, by definition, the results indicating one year worth of growth all reside along the horizontal line associated with an SGP of 50.

Focusing on two years' worth of learning (the olive-colored boxes above one year's worth of learning) we see that in Grade 3, if we take the student gains from 3rd to 5th grade and consider them as one-year gains, the mean SGP associated with those gains is in the mid to upper 60s depending upon the quintile of the student—meaning that those gains were achieved by 30% to 35% of 3rd grade students.

The result is striking, as it suggests that approximately one-third of all 3rd grade students make two or more years' worth of learning in a year. Or, said differently, among the 50% of students making at least a year's worth of growth, two-thirds of those students make two or more years' worth of growth. That is, out of 100 students, by definition 50 will attain less than a year's worth of learning, 17 will attain between one and two years of learning, and 33 will demonstrate two or more years' worth of learning.

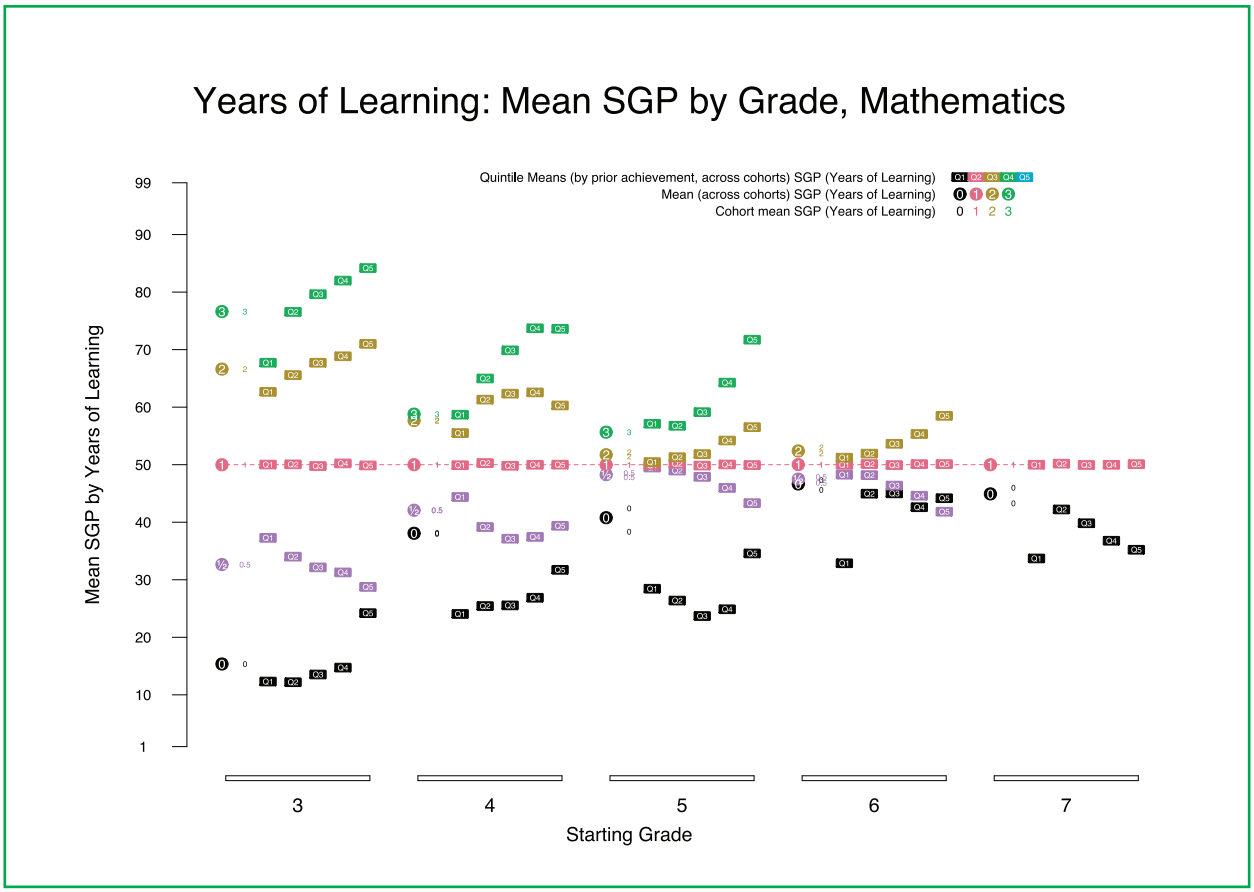


Figure 4: Mean SGP by years of learning based upon starting grade of student and achievement quintile for one state's assessment data.

Looking at zero years' worth of learning (black boxes) for grade 3, we see that the mean SGP ranges between 10 and 20 for Q1 to Q4 and is approximately 25 for Q5. Again, the results are striking: Approximately 25% of the highest achieving students have zero (or negative) learning.

Going up in grade, one can see that the percentage of students demonstrating at least two years' worth of learning become more frequent (SGP associated with exactly two years' worth of learning gets lower) and the percentage of students demonstrating no or negative learning increases as well. Results from the three other states we examined are generally consistent.

The results, taken at face value, are not believable. Their impossibility logically implies a flaw in the premise that we can infer zero or two years' worth of learning for individual students via the vertical scales we employ. Taken further, the whole enterprise of talking about months and weeks of learning is not supported by the vertical scales that we employ. As we discussed previously, on its face the conversion of a scale with questionable interval properties to one (i.e. time) with ratio properties is a non-starter. These empirical results confirm that.

A more modest interpretation

Though efforts to convert scale differences into time-based "years of learning" metrics is not supported, that doesn't imply that a more modest conversion isn't possible—one that's grounded in

empirical data and group means. For example, taking 50th percentile growth as "a year's worth of learning," it is likely defensible to produce either a dichotomy or trichotomy:

- Dichotomy: Below a year's worth of learning & at/above a year's worth of learning
- Trichotomy: Substantially below a year's worth of learning, at or near a year's worth of learning, substantially above a year's worth of learning.

Such an approach avoids the nonsense associated with weeks/months of learning while still keeping some semblance of time in place. In the next section we push these ideas further in discussing the utility of time-based interpretations of learning.

UTILITY

"I used to say, 'It is better to be complicated than wrong.' But recently I've relaxed this perspective when engaging with the public about measurement. I'd rather people be interested and learn more than tune out entirely. This comes with risks. I'll accept them." – Andrew Ho (Sept. 1, 2022)

The quote above from Andrew Ho, describing the evolution of his feelings about the conversion of score trends into "weeks of learning," is a fitting introduction to the final section of our paper, in which we discuss the utility of such time-based metrics as weeks, months, or years of learning. All reporting of test scores, whether as scaled scores, scores on a vertical scale, proficiency percentages, growth scores, or transformed into accountability ratings, comes with risks of misinterpretation and misuse.

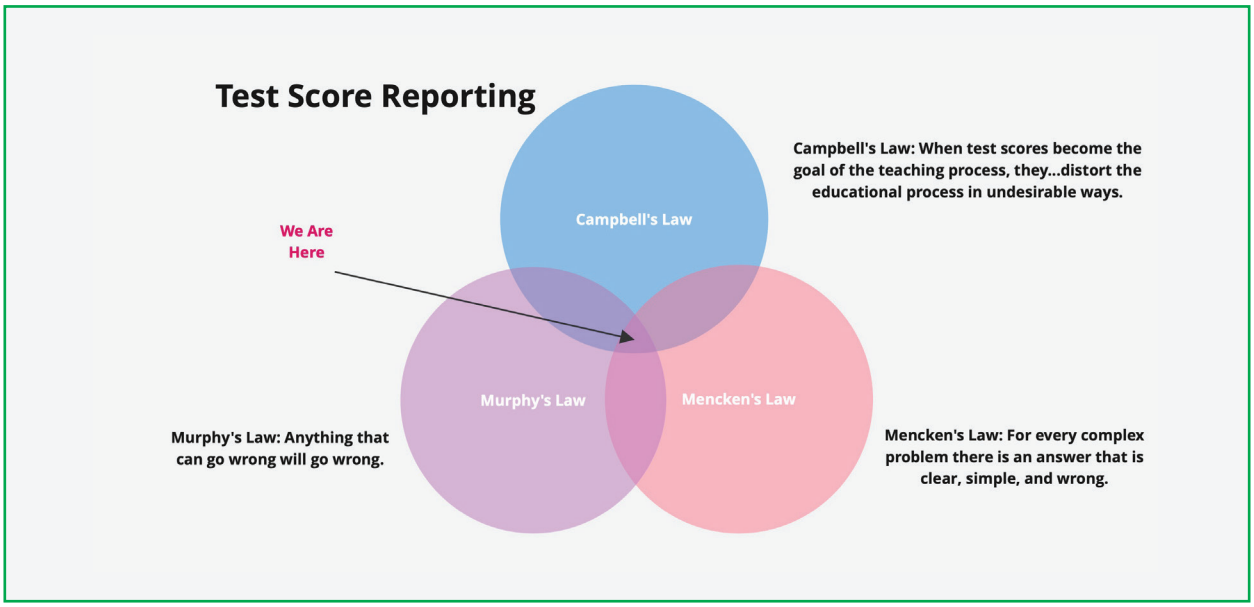
Ho's statement, made on Twitter (X) in the midst of the dual NAEP releases in the fall of 2022 (long term trend and state/main), was just one of many he made in anticipation of the specious headlines and interpretations expected to accompany what many argued, in the wake of the pandemic, was the most important release of test results in history. Other examples of so-called "misNAEPery" Ho warned about included the big three: claiming correlation is causation, psychometric misNAEPery, and one-true-outcome misNAEPery (i.e., academic achievement is the only measure/indicator that matters) (Ho, Oct 2022).

Converting effect sizes into months or weeks of learning was one of several "high crimes" and "misdemeanors" Ho cited within the category of psychometric misNAEPery, along with comparing proficiency percentages, collapsing results across subjects and grades, neglecting statistical significance, and blindly reporting differences in differences. Although all of the reporting practices above are problematic, some are more problematic than others. While acknowledging "the slippery slope from descriptive interpolation to causal inference is the primary danger of reporting score trends in 'months of learning,'" Ho considers comparing proficiency percentages without additional context (i.e., something highly likely to occur with NAEP and state test results) as the greater of two evils when compared to reporting months or weeks of learning.

Although Ho's conclusion is debatable, the fact that the reporting of test results comes down to choices between the lesser of two or more evils is, in our opinion, the bigger takeaway.

Alternative quotes considered to open this discussion of the utility of time-based conversions of test scores include cultural favorites from across the decades such as, "I'm fine. It's fine. Everything's fine," "What could possibly go wrong?" and "What, Me Worry?" All of these quotes to some degree

reflect the ongoing state of test score reporting. As a field, we readily acknowledge that we struggle to report test results to key stakeholders in a way that communicates critical information in a timely manner and in a format that is useful to them. We are not operating from a position of strength. One could argue that in a Venn diagram showing the intersection of Murphy's, Campbell's, and Mencken's Laws, test score reporting would be nesting comfortably at the intersection of the three.



There is little doubt that the idea that student achievement is two years, eight months, or six weeks *behind* generates interest. So too do headlines such as "Two Decades of Progress, Nearly Gone: National Math, Reading Scores Hit Historic Lows (Sparks, *Education Week*, 2022) and "Why 65 Percent of Fourth Graders Can't Really Read" (*The Free Press*, 2023). The risk is what follows the headlines: What happens to education policy, public perception, instruction, and student engagement? In a similar vein, we ask: To what extent does time-based reporting deepen people's understanding of student performance and/or lead to constructive actions by key stakeholders, and to what extent does it foster misinterpretations that lead to misinformed policy and bad decisions? In short: Is time-based reporting, even with all of its technical flaws, a net positive or negative?

Our position is that current efforts to convert test scores to years, months, or weeks of learning are a net negative; that is, any positive effects are far outweighed by the negative. In this section, we present three categories of criticisms of time-based reporting:

1. Time-based reporting leads stakeholders toward time-based solutions.
2. The aggregate-level results at the core of time-based reporting are not representative of what is taking place at the individual level.
3. Time-based reporting is devoid of content.

Time-based reporting leads to time-based solutions.

When we express a problem in terms of time, we're inclined to frame the solution in terms of time. We see this in the responses to descriptions of learning loss and the pandemic. When analyses of interim assessment results provided initial estimates of learning loss in terms of *how far behind*

students were, the immediate response was suggestions of ways to replace lost time (Kuhfeld et al., 2022; Cisneros, 2023; Fahle et al., 2024). Among these were recommendations for short-term or temporary solutions such as mandatory or voluntary summer school, extended school days, and tutoring, as well as calls for more permanent reforms such as extending the school year. Time-based estimates place more emphasis on recovering time lost during the pandemic than on recovering learning lost during the pandemic. The problems with such a focus are threefold.

First, time-based solutions take the focus off student learning and effective instruction. At best, solutions focused on providing additional time suggest a laissez-fare attitude toward issues related to ensuring that effective instruction to support student learning occurs during that time. At worst, they are antithetical to improving student learning, which as explained below is our second problem with time-based solutions.

Second, time-based solutions, although they may be proposed with student achievement in mind, are not focused on *increasing or improving student learning*. What do we mean by that? Recall that *student learning* is different from student achievement. It is a process, or the result of a process, that produces a change in achievement over a period of time. We can quantify student learning as the amount that achievement has changed between two points in time, which can then be expressed as the rate at which achievement changes between two points in time. Let's turn to a driving example.

Consider the scenario in which Damian and Charlie are each making the 300-mile drive from Boston to Philadelphia for the NCME conference. We can state that at a given point in time—noon for example—Charlie is an hour behind Damian. A focus solely on time would lead us to the conclusion that Charlie is going to arrive in Philadelphia an hour later than Damian. Another approach is to focus on the rate at which Charlie is traveling. Perhaps if Charlie is able to increase his velocity, he will be able to arrive in Philadelphia at the same time as Damian, or much closer to the same time. Or perhaps Charlie's velocity is such that he is losing ground at every checkpoint. If we check in again in another hour, Charlie will be even further behind.

Solutions based on adding instructional time are the equivalent of assuming that Charlie will require more time to arrive in Philadelphia. Looking more closely at Charlie's velocity is analogous to paying attention to student learning—the rate at which student achievement changes over time. It may be the case that there are factors related to the age and condition of Charlie's car and/or his body that make it impossible to increase his velocity. In such cases, additional time may be the only viable solution, but without better understanding the issues affecting his velocity and consideration of ways to increase it, we will never know.

Third, time-based solutions are not reflective of how learning occurs for individuals. The relationship between time and learning is far more complex than implied by suggestions based solely on providing additional time.

In one sense, providing additional time reflects a learning model in which a student's brain is an empty tank to be filled. In the parlance of a classic work problem, adding time in the form of extended school days, Saturday school, tutoring, etc., implies that simply adding more pipes with flowing water will fill the tank more quickly. Similarly, requiring summer school or extending the school year implies that we can simply continue to "pour knowledge" into a student nonstop, learning will occur, and achievement will increase. Both approaches ignore the possibility that there may be only so much "knowledge" that students can process in a given amount of time or without a break. We know that even in areas beset by drought, the ground can only absorb so much water at one time before it becomes saturated. The excess water is wasted, at best, and damaging, at worst.

In another sense, time-based statements of the problem and providing additional time as a solution suggest that learning and increases in achievement occur on a regular and predictable equal-interval schedule. If Damian is 6 weeks behind, after an additional 3 weeks of instruction, he will be only 3 weeks behind, in another 2 weeks, 1 week behind, and so on. In reality, increases in achievement appear to occur sporadically and in bursts. On average, and in the aggregate, student learning may appear continuous and linear, but it does not necessarily function in that manner at the individual student level, which brings us to our second category of criticisms of time-based reporting.

Aggregate-level results are not representative of what is taking place at the individual level.

The conversion of test scores to time-based estimates requires the aggregation of data collected from large samples of students—often massively large samples—over two or more points in time. The time-based estimate represents the mean, median, or some other statistically derived expected score for the sample of students selected. Under the best of circumstances, such average scores, for lack of a better word, can provide useful information to inform policy, evaluate an instructional program, or even support curricular and instructional planning for a school or classroom. Aggregate information about student achievement and growth such as average scaled scores for a school or subgroup of students, median growth scores, and even applied indices such as the percentage of proficient students in a school can provide valuable information when interpreted and used correctly. What aggregate information cannot do, even under the best of circumstances, is provide useful information about the achievement, growth, or learning of an individual student.

This reality is the conundrum that has plagued large-scale assessment since time immemorial. Large-scale assessment is neither designed to support nor intended to support or inform the instruction of individual students. Some of us have likened this situation to the difference between Newtonian physics, or classical mechanics, and quantum mechanics. We can measure an individual student's current *position* with a fairly high degree of precision, but there is a limit on the certainty with which we can discuss how they got there, where they are going next, and certainly how long it will take them to get there.

Stepping back from theoretical physics into the real world of large-scale testing, we know that there is a distribution of individual scores around any *average* score. If we know a little about the test, the sample/population of students tested, and statistics (i.e., the normal distribution), we think that we have a handle on the level of performance that a score one or two standard deviations above or below the mean represents. Time-based estimates, at least one step removed from that reality, are not easy to interpret intuitively.

As a starting point, a minor point—but an important observation—is that time-based estimates don't seem to carry with them the sense of representing the center point of a distribution in the same way that actual test scores do. Although papers published in professional journals might report effect sizes as representing a range of years, the public reporting of test scores typically does not. Reports state emphatically, and with conviction, that students are two years behind. It may be a psychological phenomenon, but we seem to be less likely to attribute distributions based on real variations in performance or errors in measurement to physical measurements than to test scores—and time is without question a physical measurement.

But we know that there is variation in any statistical estimate. We know a lot about student learning, but we understand less about how students learn. We know that the paths that individual students

follow to get from any Point A to Point B are varied, far from linear, and often not monotonic (i.e., are full of twists and turns, two steps forward followed by one step back). A cursory look at any research on learning progressions, learning maps, and the like makes this point crystal clear. So, how are we to interpret and make use of declarations such as "students are two years behind"?

We know a lot about student learning, but we understand less about how students learn.

Let's accept for the sake of argument that the time-based estimate of "two years behind" does accurately reflect how long it will take the group, on average, to reach its destination, or desired level of performance. That average two-year estimate, however, tells us nothing about how long it will take any particular individual student to reach their destination. As we stated above and demonstrated in the previous section, there is variation in students' rates of learning and in their current locations. If the average for the group is two years behind, some students may be only one year behind, not behind at all, or even ahead of where we'd expect them to be. Other students may be three, four, or more years behind. The appropriate interventions to erase that average two-year gap will differ dramatically for each of those students.

Perhaps more importantly, with regard to learning loss and recovery, we should not be interested in erasing, or making up, the **average** two-year gap. We should be interested in providing the appropriate opportunities to enable **all students** to reach the destination and reach it in a reasonable amount of time. We know how the distribution works. We can eliminate the two-year gap and return performance to where it was before the pandemic with half of the students performing above the 2019 average and half of the students performing below the 2019 average—some of them well below. That distribution, after all, is precisely how student performance looked in 2019. But normal isn't going to cut it in a new post-pandemic normal. The goal now, as it was when the No Child Left Behind Act was enacted in 2002, is for all students to reach the established and agreed upon destination, and to reach it as efficiently as possible.

Getting back to the 2019 level of student achievement is a very conservative benchmark. Had no pandemic occurred, states should have been above 2019 levels by now (in 2024). That is, percent-proficient results would almost certainly be higher today than they were in 2019 had the pandemic not occurred. Yet we compare current results to 2019 as though they are the referent of where we would have been had the pandemic not occurred.

To overcome this issue, we need concrete information about where individual students are, how far they need to go, what obstacles lie in their paths, and their rates of learning. We need to know what they can do, what they cannot do, and what they need to know and be able to do to move forward. And that brings us to the third criticism of time-based reporting.

Time-based reporting is devoid of content.

Like its Grade Equivalent (GE) score forebears, current time-based reporting of test scores is devoid of content. The estimate that a group of students is two years behind, 18 weeks behind, or has made a year's worth of recovery, by design, derivation, and definition (i.e., the way that it was estimated) provides no information about what students as a group or individually, know and are able to do. More importantly, the time-based estimate provides no information on specific gaps that need to be filled to help them reach their learning destination.

To be fair, disconnection from content is not unique to time-based reporting. It is a factor to be dealt

with in the reporting of all large-scale test results which are based on unidimensional IRT models and designed to provide estimates of a student's overall level of proficiency in the content area being assessed. Sub-scores (for better or worse), item-level statistics, and related item- or domain-based reports provide some level of content-based information, particularly at the group level, but we are not here to argue that a scaled score or achievement level classification is the gold standard for providing information about what a student knows and is able to do. The disconnection from content, however, increases significantly the further the reporting statistic is removed from the content of the test.

Growth Scores

Growth scores, describing a student's change in performance between two test administrations, provide less content-related information than scaled scores. The lack of interpretable content information is easily illustrated when growth scores are based on a change in student performance on a vertical scale. As discussed previously with regard to GE scores, it is not possible to make content-based interpretations of two scores on a vertical scale when those scores are based on student performance on different tests (and hence different base scales). Any attempt at making a content-based interpretation requires knowledge of which test the student was administered. Again, as discussed in the section on GE scores, interpretation in general, and content-based interpretation, in particular, is particularly problematic for students performing below grade level.

Time-based reporting is one step further removed from the content of a particular test.

Time-based reporting of test scores, which often depends heavily on statistical manipulation of reporting scales, is inherently detached from the actual content of tests. Such reports are based on assumptions about standard deviations and do not provide concrete content-related insights. Moreover, translating time-based metrics (e.g., being "two years behind") back into meaningful, content-specific interpretations is impractical without additional detailed analyses that could have been directly conducted initially, bypassing the need for a time-based approach altogether. Andrew Ho's observation underscores this point, suggesting that while time-based reports come with some risks, they might prompt policymakers to request further analysis. This potential benefit, however, is often outweighed by the risk of misinterpretation for both groups and individual students.

Understanding what students know and can do is crucial for accurately diagnosing educational challenges and effectively planning interventions. This is analogous to understanding traffic flow in a jam: knowing that a 10-mile trip takes 90 minutes offers limited insight without identifying specific bottlenecks that slow progress. Allowing myself 90 minutes to make my commute (i.e., adding extra time) might get me to work on time (i.e., achieve the desired outcome), but does nothing to improve my commute (the process). Tools like Google Maps offer alternate routes that are more efficient; leaving the house just 10 minutes earlier or 15 minutes later might cut my travel time in half. The Maps app on my iPhone provides detailed, color-coded (blue, yellow, red) information that identifies specific problem areas in my commute that might benefit from targeted interventions.

CONCLUSIONS

Despite the case that we have made in this paper against time-based reporting, it is true that the solution to recovery from the pandemic, closing achievement gaps, and education reform in general, involves improving student learning, and student learning is inextricably linked to time. Improvement, however, is based on the **effective use of time**, not simply on the addition of it.

Improvement is based on the **effective use of time**, not simply on the addition of it.

When we attempt to determine the amount that student achievement changes over the course of a year (define a "year of learning"), our goal is to better understand how much "learning" occurs in a year under current conditions, with the ultimate goal of increasing that amount, and doing so by **improving the rate** of student learning. We conclude that the rate of student learning has improved when we see 1) an increase the amount that a student's achievement changes in a fixed amount of time or 2) a decrease in the amount of time that it takes for students to attain a fixed change in achievement status.

It is the combination of time and achievement (change in achievement over time) that defines students learning.

Either way, it is the combination of time and achievement (change in achievement over time) that defines student learning. We need to increase the rate of student learning by making the curriculum more engaging, making instruction more effective, or by reducing any, some, or all of the myriad barriers that have made the task of improving student learning seem insurmountable for so many years. The instinct to focus reporting on time, therefore, is not totally misguided; it is simply an incomplete representation of the construct.

REFERENCES

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., ... & Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. EPI Briefing Paper# 278. *Economic Policy Institute.*

Bennett, R. E. (1982). The use of grade and age equivalent scores in educational assessment. *Diagnostique, 7*(3), 139-146.

Berk, R. A. (1981). What's wrong with using grade-equivalent scores to identify LD children? *Academic Therapy, 17*(2), 133-140.

Bryant, J., Dorn, E., Pollack, L., & Sarakatsannis, J. (2023). COVID-19 learning delay and recovery: Where do US states stand? McKinsey & Company. Downloaded from https://www.mckinsey.com/industries/education/our-insights/covid-19-learning-delay-and-recovery-where-do-us-states-stand

Carnegie Foundation (2023). Carnegie Foundation, ETS partner to transform the educational pillars they built: The Carnegie Unit and standardized tests. Downloaded from: https://www.carnegiefoundation.org/newsroom/news-releases/carnegie-foundation-ets-partner-to-transform-the-educational-pillars-they-built-the-carnegie-unit-and-standardized-tests/

Cisneros, J. (2022). Should the school year be extended to make up for learning loss? NEXTSTAR. Downloaded from https://cepr.harvard.edu/news/should-school-year-be-extended-make-learning-loss

Curriculum Associates (2020). Understanding student needs: Early results from fall assessments. Curriculum Associates Research Brief. Downloaded from https://www.curriculumassociates.com/-/media/mainsite/files/i-ready/iready-diagnostic-results-understanding-student-needs-paper-2020.pdf

Fahle, E., Kane, T. J., Reardon, S. F., & Staiger, D. O. (2024). The first year of pandemic recovery: A district-level analysis. Education Recovery Scorecard. Center for Education Policy Research: Harvard University. The Educational Opportunity at Stanford University. Stanford cepa. Downloaded from https://educationrecoveryscorecard.org

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145-220.

Ho, A. (Sept 1, 2022). Discussion of long term NAEP results. Twitter thread. Downloaded from: https://twitter.com/AndrewDeanHo/status/1565345522050691077

Ho, A. (Oct 22, 2022). Let's talk misNAEPery. Twitter thread. Downloaded from: https://twitter.com/AndrewDeanHo/status/1583939513822638081

Kerns, G. (2023). AssessMinutes – Understanding grade equivalent scores. Renaissance. Viewed at https://youtu.be/IpIUFj9fuUs?si=85sMvDdqIX0-Qpsg

Kuhfeld, M., Soland, J., Tarasawa, A.J., Ruzek, E., & Liu, J. (2020). Projecting the potential impacts of COVID-19 school closures on academic achievement. Annenberg EdWorkingPapers. Downloaded from https://edworkingpapers.com/ai20-226

Kuhfeld, M., Soland, J., Lewis, K., & Morton, E. (2022). The pandemic has had devastating impacts on learning. What will it take to help students catch up? Brookings. The Brown Center Chalkboard. Downloaded from https://www.brookings.edu/articles/the-pandemic-has-had-devastating-impacts-on-learning-what-will-it-take-to-help-students-catch-up/

Kuhfeld, M. Dilliberti, M. McEachin, A., Schweig, J., & Mariano, L.T. (2023). Typical learning for whom? Guidelines for selecting benchmarks to calculate months of learning. NWEA Research. Downloaded from https://www.nwea.org/uploads/Guidelines-for-selecting-benchmarks-to-calculate-months-of-learning_NWEA_Research-Brief.pdf

Merod, A. (2023). Students need over 4 months of extra learning to return to pre-pandemic math, reading achievement. K-12 Dive Brief. Downloaded from https://www.k12dive.com/news/learning-loss-recovery-research-NWEA/686153/

Patrinos, H.A., Vegas, E., & Carter-Rau, R. (2022). An analysis of COVID-19 student learning loss. Education Global Practice for World Bank Group. Downloaded from https://documents1.worldbank.org/curated/en/099720405042223104/pdf/IDU00f3f0ca808cde0497e0b88c01fa07f15bef0.pdf

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational leadership, 56*, 8-16.

Ramos, C. (1996). The computation, interpretation, and limits of grade equivalent scores. https://libcat.colorado.edu/Record/b6392873

Renaissance (2020). How kids are performing: Tracking the impact of COVID-19 on reading and mathematics achievement. Fall 2020 Edition. Special Report Series. Downloaded from https://renaissance.widen.net/view/pdf/nrkoqeyesg/R63289.pdf?u=zceria&t.download=true

Schwartz, S. (2021). Learning loss in general is a misnomer. Study shows kids made progress during COVID-19. *Education Week.* April 2021. Downloaded from https://www.edweek.org/leadership/learning-loss-in-general-is-a-misnomer-study-shows-kids-made-progress-during-covid-19/2021/04

Sparks, S.D. (2022). Two decades of progress, nearly gone: National math, reading scores hit historic lows. *Education Week.* 10/24/2022. Downloaded from https://www.edweek.org/leadership/two-decades-of-progress-nearly-gone-national-math-reading-scores-hit-historic-lows/2022/10

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680.

The Free Press (2023). Why 65 percent of fourth graders can't really read. 2/11/2023. Downloaded from https://www.thefp.com/p/why-65-percent-of-fourth-graders

UnboundEd (2021). Why unfinished instruction is more accurate and equitable than learning loss. Downloaded from https://unbounded.org/resources/why-unfinished-instruction-is-more-accurate-and-equitable-than-learning-loss/

Whitby. T, Thomas, S. & Wirtz, R. (2021). Learning loss: A real concern, a deficit mindset or an overblown debate? EdChat Radio Podcast. Downloaded from https://www.bamradionetwork.com/track/learning-loss-a-real-concern-a-deficit-mindset-or-an-overblown-semantic-debate/

Wormeli, R. (2022). Differentiated instruction: Principles, myth busting, & practicalities. Downloaded from: https://ksdetasn.s3.amazonaws.com/uploads/resource/upload/3318/Kansas_MTSS_January_2022_Session_1_PDF_Differentiated_Instruction_Principles_Myth_Busting_Practicalities_handout_version.pdf

Zenisky, A. L., & Hambleton, R. K. (2015). A model and good practices for score reporting. In *Handbook of Test Development* (pp. 585-602). Routledge.

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment, 19*(2), 116-138.

National Center for the Improvement
of Educational Assessment
Dover, New Hampshire

www.nciea.org